

Information Retrieval

Theodore Kalamboukis, Professor AUEB, tzk@aueb.gr

Overview

Word statistics, Vector space model (relevance feedback, query expansion, document normalization, document reranking), evaluation of retrieval, generalized VSM, latent semantic indexing, Web retrieval, data fusion, metasearch, multimodal retrieval, applications.

Objectives and Outcomes

The main objective of this course is to present the basic concepts in information retrieval and more advance techniques of multimodal based information systems.

The second objective goal of the course is for the students to

- understand the underlined problems related to IR and
- acquired the necessary experience to design, and implement real applications using Information Retrieval systems.

Requirements and Prerequisites

The course requires basic knowledge in linear algebra, and programming skills in Java, or Python and Matlab.

Books

C.D. Manning, P. Raghavan, H. Schütze. Introduction to Information Retrieval, Cambridge UP, 2008. (available in the Web, <http://nlp.stanford.edu/IR-book/>),

Lecture-slides and other course materials will be provided.

Grading

Students will be graded on their performance as follows:

- Homework assignment 20%.
- Programming project 30%.
- Final written exams 50%.

Course Syllabus

The course comprises five units of three hours each.

Unit 1: Vector Space Model of retrieval

Word statistics, Text preprocessing, Term weighting, Similarity function, Indexing, Relevance feedback, Query expansion (with local analysis – from external resources), the impact of document normalization, Multi-field retrieval, Evaluation of retrieval. Applications.

Unit 2: Latent Semantic Indexing

Basic concepts, Singular Value Decomposition, Latent semantic indexing (LSI), LSI search engine, updating, Toward a theoretical foundation-Probabilistic analysis of LSI, applications of LSI.

Unit 3: Web Retrieval

Crawling the Web, Link analysis, Importance of ranking, Query-independent ranking (PageRank), Query-dependent ranking (HITS algorithm), personalized PageRank, applications of PageRank algorithm.

Unit 4: Data Fusion - Metasearch

Data fusion, early and late fusion, Metasearch engines of retrieval, how they work.

Unit 5: Multimedia retrieval

Multimodal retrieval, retrieving images from textual and visual data.

Unit 6: Applications

Hands on experience on real applications' data.